Séminaire de Probabilités et Statistique

Mardi 25 octobre à 14h00

Salle de conférences

Hans Kersting

Inria Paris

Exploration and Implicit Bias due to SGD's Stochasticity

The data sets used to train modern machine-learning models are often huge, e.g. millions of images. This makes it too expensive to compute the true gradient over all data sets. In each gradient descent (GD) step, a stochastic gradient is thus computed over a subset ("mini-batch") of data. The resulting stochastic gradient descent (SGD) algorithm, and its variants, is the main workhorse of modern machine learning. Until recently, most machine-learning researchers would have preferred to use GD, if they could, and considered SGD only as a fast approximation to GD. But new research suggests that the stochasticity in SGD is part of the reason why SGD works so well. In this talk, we investigate multiple theories on the advantages of the noise in SGD, including better generalization in flatter minima ('implicit bias') and faster escapes from difficult parts of the landscapes (such as saddle points and local minima). We highlight how correlating noise can help optimization and zoom in on the question which noise structure would be optimal for SGD. As a result, we propose Anti-PGD as an algorithm which perturbs GD by anti-correlated perturbations.